Exam PA June 17, 2020 Project Solution

Instructions to Candidates: Please remember to avoid using your own name within this document or when naming your file. There is no limit on page count.

Also be sure all the documents you are working on have June 17 attached.

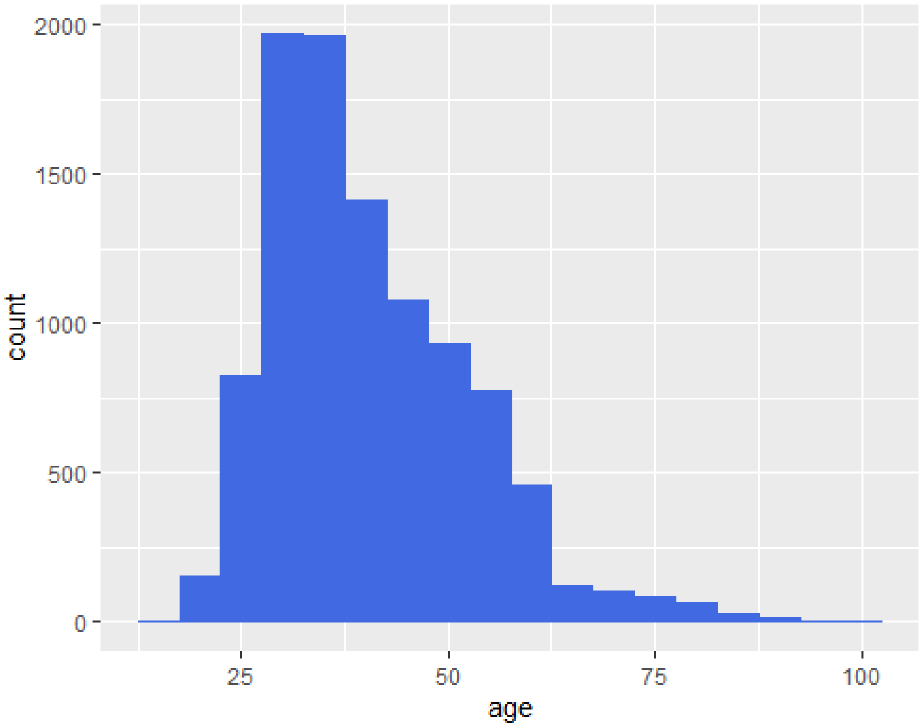
As indicated in the instructions, work on each task should be presented in the designated section for that task.

This model solution is provided so that candidates may better prepare for future sittings of Exam PA. It includes both a sample solution, in plain text, and commentary from those grading the exam, in italics. In many cases there is a range of fully satisfactory approaches. This solution presents one such approach, with commentary on some alternatives, but there are valid alternatives not discussed here.

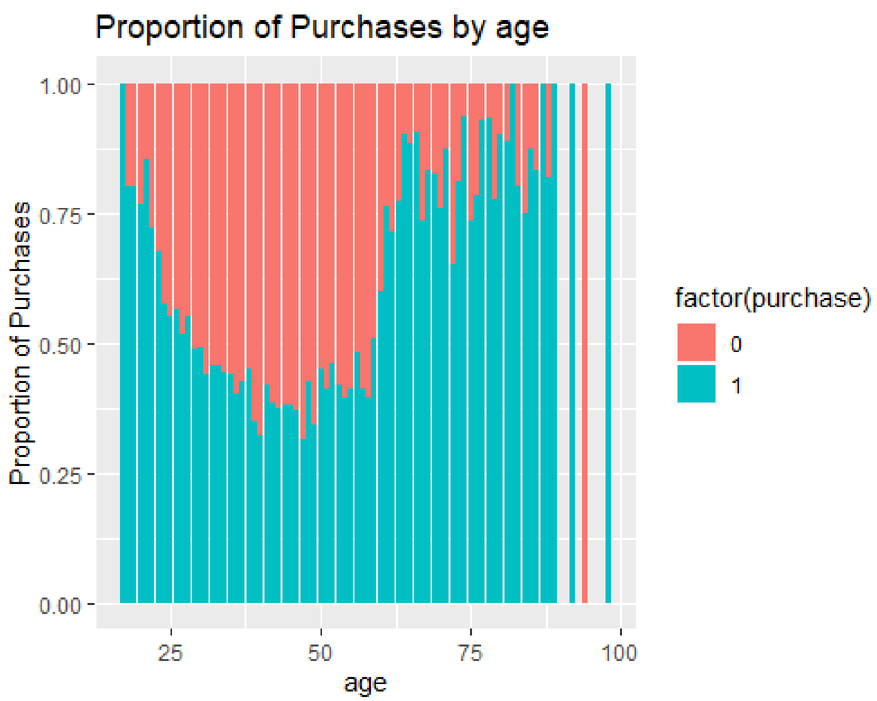
Task 1 – Explore the data (8 points)

Candidates were expected to analyze and comment on a variety of charts, but some candidates did not comment on all of them, losing credit. Many candidates did not add a comment about how it would impact future modeling but only observed how it would look statistically. That was the biggest challenge for candidates on this task.

Age

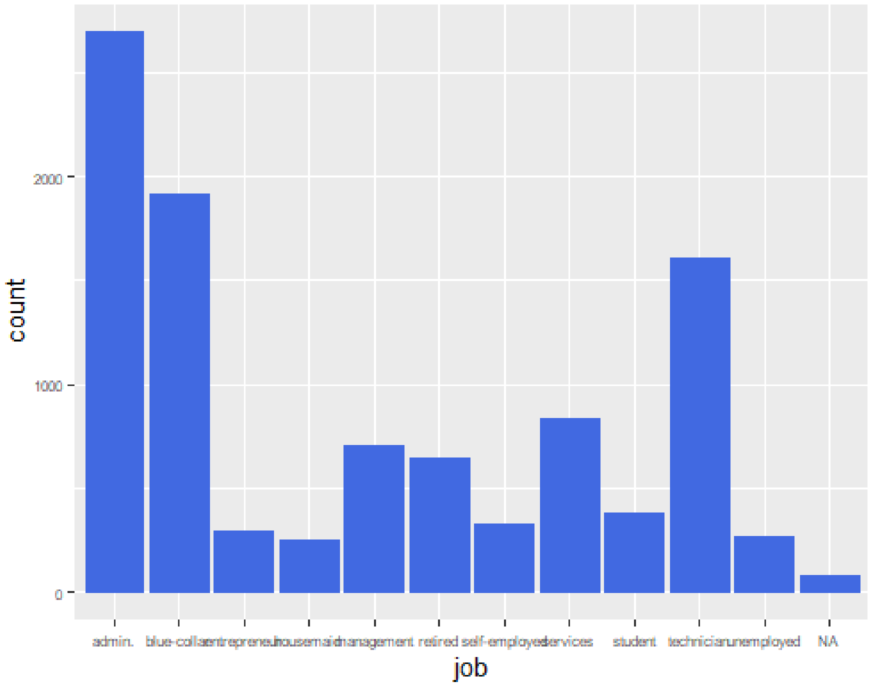


The bulk of ages are between 25 and 60, with a peak near 30. The age distribution is somewhat skewed to the right. A log transformation of age may produce a better fitting model and should be considered along with no transformation of age.

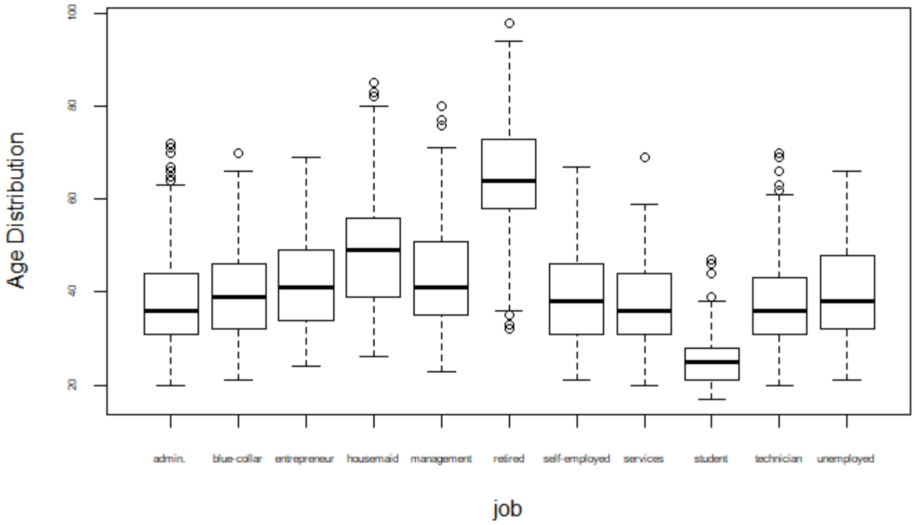


The proportion of purchase has a downward trend roughly between age 17 and age 50 and then starts increasing after that, particularly a significant jump around age 60 and thereafter. GLM models will have trouble fitting this down-up curve in purchase by age with only a single age variable, and additional variables based on age, for example age squared, will be needed to capture the observed trends for GLM. Decision trees will not need additional variables to capture such shapes.

Job

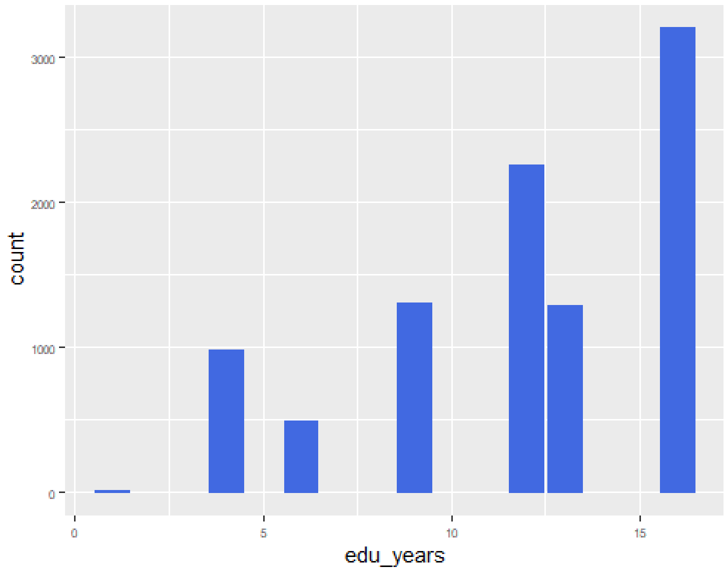


The highest proportion of ABC prospective customers has an administrative job, followed by blue-collar and technician jobs. Those missing a job categorization are rather small, but their relative impact on purchase rate should be noted before deciding how to handle these missing values. There are many categories, which may lead to high variance when fitting the model. If ABC could help to combine some of the categories, the resulting models may produce more reliable predictions based on job.

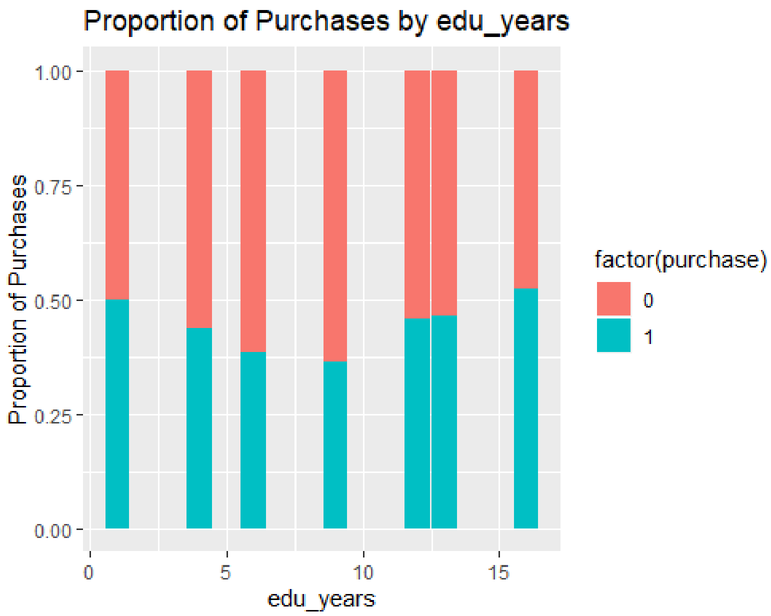


In looking at age by job, two categories, retired and student, pop out immediately. It makes sense that retirees tend to be at older age while students are more likely to be at younger age. As age and job have some codependence, I need to be careful in dealing with these two variables together in the same model. Particularly for greedy decision trees but also possible for GLM, strong results for one variable may hide the influence of the other variable.

Edu_Years



Edu_years takes integer values from 1 to 16, with noticeable gaps between values. It may be difficult to decide how to assign missing values given this distribution. Also, there are very few observations having a value of 1, an outlier that may need to be removed given its distance from other observations.



For the proportion of purchases by edu_years, there is a bowed shape, having higher proportions of purchase near the two end points but lower proportions of purchase in the middle. A linear model may miss this due to much higher frequencies of higher edu_years. Because of this, I need to consider treating it is a factor variable for a GLM or using models like tree-based models that can handle non-linear patterns well.

Task 2 – Consider the education variable (3 points)

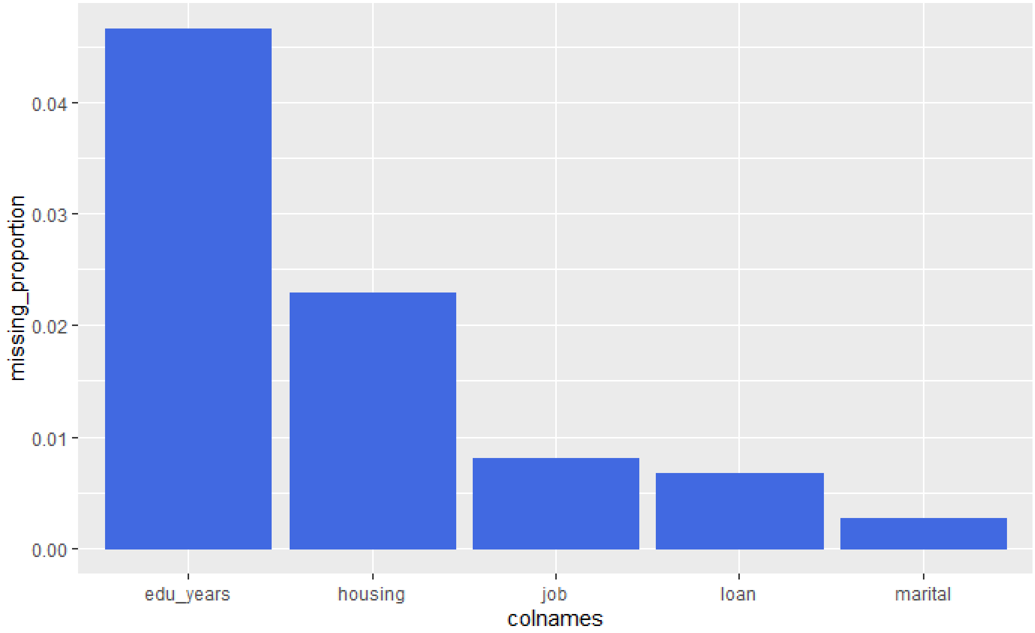
> *Few candidates considered the difference in dimensionality from using a numeric variable. Often, candidates did not distinguish between GLM and decision trees. Some candidates, when discussing decision trees, compared numerical and categorical and said one was better than another without recognizing that, in this case, the decision tree could ultimately reproduce the original categories.*

A strategy for dimension reduction is necessary to avoid the curse of dimensionality, which can lead to overfitting. The original categorical variable on education requires six variables, in addition to a baseline category, in the model. On the other hand, creating a new numeric variable, edu_years, only requires one variable. While some information is lost in going from six variables to one, the risk of overfitting can be greatly reduced depending on the model used.

GLM models can better discern linear trends over several ordinal categorical variables, as seen in edu_years, when they are converted to numerical variables. For decision trees, however, the conversion to numerical does not reduce the dimensionality as much because the tree can still split between any adjacent pair of variables and, with enough splits, reproduce the categorical variables.

Task 3 – Handle missing values (5 points)

The edu_years variable presents a challenge as each method for removing missing data has material flaws. Overall, alternative approaches for variables could often be justified.



```
[1] "Purchase Proportions by variable, for missing and non missing values"
[1] "  Variable      PP_for_NAs  PP_for_non_NAs"
[1] "   housing            0.47             0.46"
[1] "       job            0.46             0.46"
[1] "      loan            0.52             0.46"
[1] "   marital            0.44             0.46"
[1] " edu_years            0.54             0.46"
```

Edu_years: impute using mean. Almost 5% are missing, and the purchase proportion is significantly higher for missing values than non-missing values. Removing these may cause us to lose valuable insights. Being a numeric variable, converting to an “unknown” value does not work, so imputing the missing value using the mean is left, though I do not have confidence that what causes these to be missing is spread evenly among education.

Housing: convert to “unknown”. As over 2% are missing, converting to unknown despite not much difference in purchase proportions between missing and non-missing values.

Job: remove rows. Less than 1% are missing and the purchase proportion is almost identical to that of non-missing values.

Loan: convert to “unknown”. Less than 1% are missing, but the purchase proportion of these is noticeably different to that of non-missing values, so the missingness might be predictive.

Marital: remove rows. Less than 0.5% are missing and the purchase proportion is similar to non-missing values.

Task 4 – Investigate correlations (3 points)

Which correlations are concerning can be a matter of judgment, though the 94% correlation between irate and employment clearly needed to be discussed. Most candidates did not relate concerns on the correlations to specific modeling techniques, which was needed to earn full credit. Some candidates mentioned clustering as an alternative technique, but the type of clustering was important, as some clustering approaches do not easily accommodate new observations for prediction.

```
                   age   edu_years         CPI         CCI       irate  employment
age         1.00000000 -0.23990642 -0.01632691  0.14089152 -0.04561588 -0.07612409
edu_years  -0.23990642  1.00000000 -0.09182075  0.03748592 -0.08288903 -0.08025666
CPI        -0.01632691 -0.09182075  1.00000000 -0.14953031  0.57589245  0.35507393
CCI         0.14089152  0.03748592 -0.14953031  1.00000000  0.05887682 -0.07287912
irate      -0.04561588 -0.08288903  0.57589245  0.05887682  1.00000000  0.94114991
employment -0.07612409 -0.08025666  0.35507393 -0.07287912  0.94114991  1.00000000
```

The most notable correlations are:

- irate and employment (0.94)
- CPI and irate (0.58)

These correlations and others are not that concerning for decision trees. For example, irate and employment are heavily correlated, and so no or little information can be gained from splitting on employment after having split on irate and the second variable will be excluded. The only concern is that the variable chosen may flip-flop depending on training data and the modeler may not be aware that the other is almost as predictive.

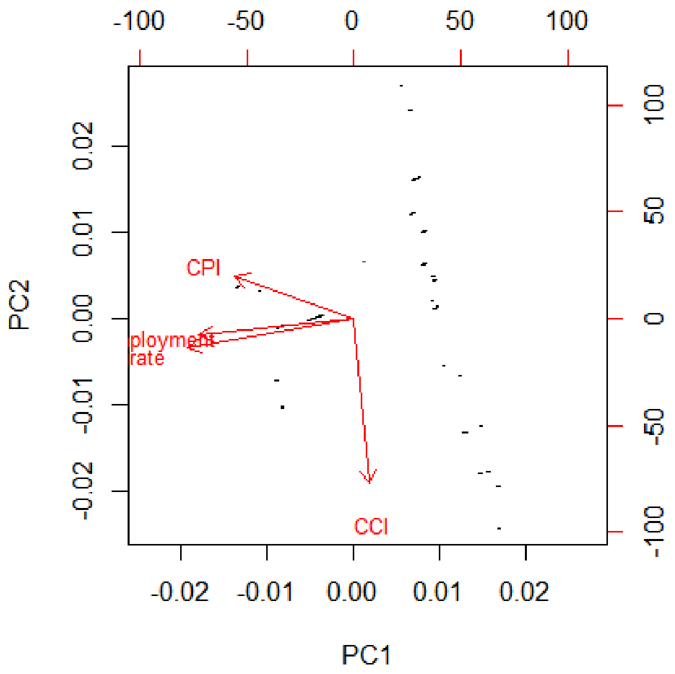
These correlations are concerning for GLM models, which do not handle highly collinear variables well. Very large and mostly offsetting coefficients may result, making interpretation of the coefficients difficult. In particular, it is dangerous to interpret the coefficient as representing the impact on the target variable with other variables held constant, given that the correlated variable is likely to also change. The accuracy of the estimated coefficients is also questionable and different results can occur if a new sample is taken.

One method other than PCA for handling the correlated variables is to use one of the variables and delete the redundant ones.

Task 5 – Conduct a principal components analysis (8 points)

Candidates generally printed the bi-plot but often did not indicate how to read the plot when explaining the loadings. Few candidates interpreted the plot very well, but many candidates used the proportion of variance well when choosing how many components to include. Few candidates sufficiently addressed why scale will affect the PCA results.

A good way of handling correlated variables is to perform principle components analysis (PCA) to obtain orthogonal variables in which different principle components are uncorrelated, but still containing most of the information. Here, I did a PCA on the following variables: CPI, CCI, irate, and employment. I set the scale parameter to “TRUE” so variables can be scaled to have unit variance. Without it, certain variables could dominate the associations between the variables due to larger magnitudes of variance.



In the PCA bi-plot, the relative loadings as seen in the red scales and arrows are of the most interest for comparing the first two principle components. Employment and irate have nearly identical positions, showing that PC1 and PC2 do not distinguish much between them. In PC1, similar movements in these two variables and CPI are grouped together with little emphasis on CCI, while PC2 highlights movements in CCI, combined with some opposing movement in CPI. The variation from PC2 is visible in the black PC scores, with a wide (tall, really) variation of PC2 scores for PC1 scores between 0.01 and 0.02.

```
        PC1         PC2         PC3         PC4
0.571608387 0.260337336 0.165255043 0.002799234
```

The proportion of variance explained by each succeeding principal component is noted above. I recommend using **two** principal components in the GLM model, because PC1 and PC2 explain 83% of the variation and each can be interpreted in a straightforward manner based on the original variables.

Task 6 – Create a generalized linear model (5 points)

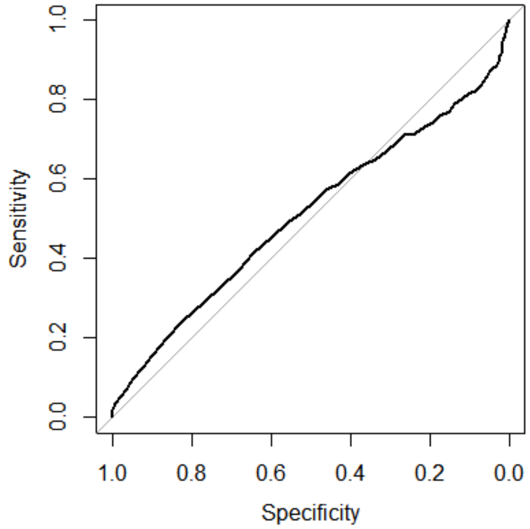
> *While most candidates had little trouble with the output of the model, some candidates struggled with the explanation on the differing performance of the age variable. The ROC curves and AUC were given as a convenience to help recognize the difference in model performance, but other ways could have been used to compare the models.*

To begin, an age-nly GLM with a logit link was built on the training data set.

```
Call:
glm(formula = purchase ~ age, family = binomial(link = "logit"),
    data = data_train)

Deviance Residuals:
   Min       1Q   Median       3Q      Max
-1.268   -1.116   -1.084    1.243    1.303

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.399993   0.085297  -4.689 2.74e-06 ***
```

```
age          0.006490   0.002027   3.201  0.00137 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9571.2  on 6927  degrees of freedom
Residual deviance: 9561.0  on 6926  degrees of freedom
AIC: 9565
```

Based on the output shown above, the age variable has a low p-value, showing that, in isolation, it is a statistically significant predictor.

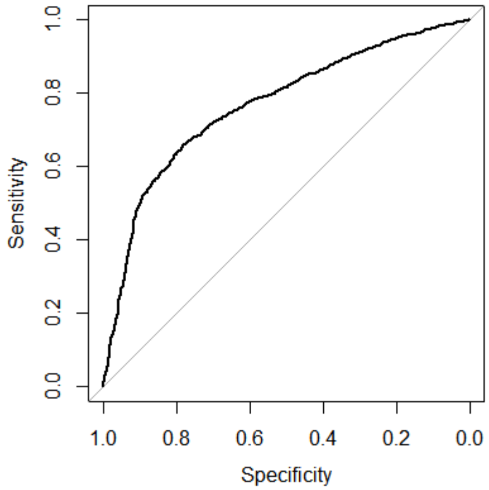


In addition, the area under the above ROC curve (AUC) for the test data is 0.5099. Clearly, this age only single factor model is not doing very well. Its performance is little different from the 0.5 expected for an intercept-only model.

Next, I ran a full GLM model with a logit link on the following variables.

- age
- job
- marital
- edu_years
- housing
- loan
- phone
- month
- weekday
- PC1
- PC2

```
Call:
glm(formula = purchase ~ age + job + marital + edu_years + housing +
    loan + phone + month + weekday + PC1 + PC2, family = binomial(link = 
"logit"), 
    data = data_train)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.4201  -0.8656  -0.5140   0.8511   2.2113  

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)    
(Intercept)      -0.449525   0.264086  -1.702 0.088719 .  
age               0.004356   0.003322   1.311 0.189716    
jobblue-collar   -0.226631   0.100820  -2.248 0.024584 *  
jobentrepreneur   0.055879   0.161880   0.345 0.729952    
jobhousemaid     -0.292176   0.197269  -1.481 0.138579    
jobmanagement    -0.094480   0.116912  -0.808 0.419015    
jobretired        0.252304   0.155862   1.619 0.105497    
jobself-employed -0.183896   0.159940  -1.150 0.250233    
jobservices      -0.060748   0.113645  -0.535 0.592965    
jobstudent        0.445711   0.169717   2.626 0.008634 ** 
jobtechnician     0.032976   0.086026   0.383 0.701476    
jobunemployed     0.188435   0.184660   1.020 0.307520    
maritalmarried    0.026409   0.092808   0.285 0.775984    
maritalsingle     0.106392   0.106281   1.001 0.316803    
edu_years         0.012444   0.009758   1.275 0.202256    
housingyes       -0.100109   0.057041  -1.755 0.079254 .  
housingunknown   -0.141871   0.226197  -0.627 0.530528    
loanyes          -0.091265   0.076934  -1.186 0.235512    
loanunknown       0.452525   0.441468   1.025 0.305342    
phonelandline    -0.082722   0.091431  -0.905 0.365597    
monthaug          0.021748   0.147905   0.147 0.883102    
monthdec          1.221970   0.482771   2.531 0.011369 *  
monthjul          0.473005   0.128605   3.678 0.000235 ***
monthjun          0.525706   0.130077   4.041 5.31e-05 ***
monthmar          0.887184   0.205342   4.321 1.56e-05 ***
monthmay         -0.677697   0.106493  -6.364 1.97e-10 ***
monthnov         -0.189179   0.131791  -1.435 0.151160    
monthoct          1.040555   0.211429   4.922 8.59e-07 ***
monthsep          0.657589   0.223044   2.948 0.003196 ** 
weekdaymon       -0.147597   0.090045  -1.639 0.101183    
weekdaythu        0.103372   0.088609   1.167 0.243368    
weekdaytue        0.003959   0.090481   0.044 0.965103    
weekdaywed        0.122229   0.089942   1.359 0.174158    
PC1               0.627463   0.027530  22.792  < 2e-16 ***
PC2               0.039072   0.039754   0.983 0.325689    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9571.2  on 6927  degrees of freedom
Residual deviance: 7716.7  on 6893  degrees of freedom
AIC: 7786.7
```



This model is performing much better with a test data AUC of 0.7665, showing that the model makes some effective predictions due to the additions of many other predictors. Given the much better AUC, it does not feel like age is among the more robust predictors. Age now has a high p-value and is no longer in itself a good predictor. Age is not independent from other variables, particularly job, and the trends ascribed to age in the age-only model are much better described by other variables, for instance "job: student", once they are included in the model.

Task 7 – Select features using stepwise selection (8 points)

Most candidates who answered all parts of the question did well. However, many candidates skipped the discussion on best subset selection—those who answered it mentioned efficiency, appropriately. Fewer noted how stepwise processes may not find the best model.

Many candidates did not explicitly list the variables chosen as asked for in the question.

Any combination of stepwise process choices could earn full credit when well justified.

For the purpose of feature and model selection, one can use best subset selection, fitting separate GLMs for each possible combination of features and then select the best combination. However, when there are many predictors in a model, the number of possible combinations can be very large, making best subset selection impractical and computationally inefficient.

An alternative is stepwise process, constructing a model by adding (forward selection) or removing (backward selection) one predictor at a time. It is a simpler and faster process compared to best subset selection, but it may not find the optimal combination of features.

With forward selection the model starts with no variables and then adds one variable at a time until there is no improvement in the selected criterion. Backward selection starts with all the variables and sequentially removes them until no improvement results. It is more likely that forward selection will result in a simpler model.

When fitting models by maximum likelihood, additional variables never decrease the loglikelihood. For AIC, adding a variable requires an increase in the loglikelihood of two times the number of parameters added. For BIC, the required per parameter increase is the logarithm of the number of observations. BIC is a more conservative approach as there is a greater penalty for each parameter added, requiring more evidence to support additional variables.

Since our goal in this project is to identify the key variables that relate to the target variable, it makes sense to take a conservative approach. Thus, forward selection is chosen. BIC was originally selected for this analysis. However, it ended up with a model that has two variables, PC1 and month. This is too few to be helpful for ABC, so AIC is selected for this analysis to provide more useful results. Therefore, I select the combination of forward selection and AIC.

The summary report of the resulting model can be found below.

```
Call:
glm(formula = purchase ~ PC1 + month + weekday + job, family = binomial(link
= "logit"),
    data = data_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4295  -0.8625  -0.5331   0.8478   2.1372

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.107928   0.117437  -0.919 0.358083
PC1               0.644329   0.023459  27.466  < 2e-16 ***
monthaug         -0.029188   0.117221  -0.249 0.803357
monthdec          1.120358   0.474668   2.360 0.018260 *
monthjul          0.469336   0.121083   3.876 0.000106 ***
monthjun          0.483871   0.126402   3.828 0.000129 ***
monthmar          0.873300   0.204628   4.268 1.97e-05 ***
monthmay         -0.721704   0.103431  -6.978 3.00e-12 ***
monthnov         -0.221300   0.123732  -1.789 0.073688 .
monthoct          0.994786   0.207339   4.798 1.60e-06 ***
monthsep          0.588893   0.215219   2.736 0.006214 **
weekdaymon       -0.142614   0.089897  -1.586 0.112645
weekdaythu        0.103931   0.088419   1.175 0.239822
weekdaytue        0.001999   0.090362   0.022 0.982353
weekdaywed        0.121876   0.089721   1.358 0.174344
jobblue-collar   -0.295904   0.084227  -3.513 0.000443 ***
jobentrepreneur   0.029427   0.160386   0.183 0.854422
jobhousemaid     -0.345778   0.187038  -1.849 0.064502 .
jobmanagement    -0.078057   0.115325  -0.677 0.498504
jobretired        0.267538   0.129859   2.060 0.039377 *
jobself-employed -0.181231   0.159531  -1.136 0.255948
jobservices      -0.102466   0.110348  -0.929 0.353108
jobstudent        0.392211   0.161797   2.424 0.015347 *
jobtechnician     0.022490   0.085542   0.263 0.792615
jobunemployed     0.144544   0.182462   0.792 0.428253
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9571.2  on 6927  degrees of freedom
Residual deviance: 7728.9  on 6903  degrees of freedom
AIC: 7778.9
```

The final model from stepwise selection has picked the following variables:

- PC1
- Month
- Weekday
- Job

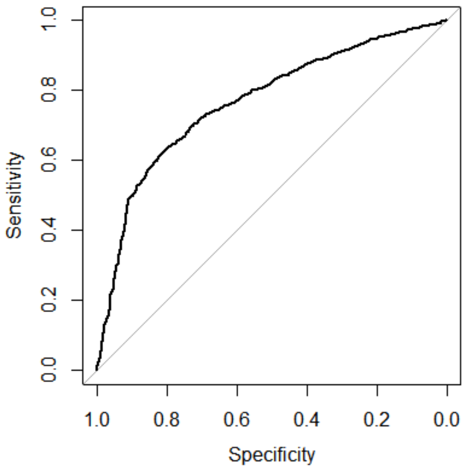
Task 8 – Evaluate the model (12 points)

On AUC, most candidates described AUC near 1 well but many struggled with explaining 0.5 (which is not just half right and half wrong) and near 0. Typically, an AUC of less than 0.5 would not happen and would indicate a problem in the model optimization or metric calculation.

Many candidates did not remember to compare variables selected to previous tasks as stated in the problem statement.

Many candidates did well to provide observations to marketing beyond numerical interpretation, making them more discussable and actionable.

The final model from Task 7 has AUC's of 0.7780 and 0.7669 on training and testing, respectively, indicating the model fit is good overall. The ROC curve for the test data is shown below.



A perfect model that predicts the correct class for new data each time will have a ROC plot showing the curve approaching the top left corner with an AUC near 1.0. When a model has an AUC of 0.5, like when the ROC curve runs along the diagonal shown, its performance is no better than randomly selecting the class for new data such that the proportions of each class matches that of the data. Any model having an AUC less than 0.5 means it is providing predictions that are worse than random selection, with a near 0 AUC indicating that the model makes the wrong classification almost every time.

In the data exploration, age and month were expected to have an impact on the proportion of purchase, but the other models found age was not (linearly) predictive while month looks impactful in all models. Compared to the second GLM in Task 6, only 4 of the 11 variables are retained after feature selection was applied, but the dropped variables had looked insignificant in that model's results.

The logit function is the natural log of odds, log(p/(1-p)), where p the probability of purchase and p/(1-p) defines the odds of purchase. Thus, the correct way of interpreting coefficients is to exponentiate them, providing the odds factor for a given predictor. The table below summarizes this interpretation for select features.

Feature	Coefficient	Interpretation
PC1	0.644	A one-unit change in PC1 results in increasing odds of purchase by exp(0.644)=190%. This better chance for purchases corresponds to lower employment, interest rates, and CPI.
monthmar	0.873	The odds of purchase in March is 240% times that of the baseline month, April. Interpretations for other months are similar in this fashion. The highest month (where January and February are not available in the data) is December; the lowest month is May.
weekdaymon	-0.143	The odds of purchase on Monday is 87% times that of the baseline day, Friday, which indicates the likelihood of purchasing on Monday is lower than the likelihood of buying on Friday, relatively speaking. Monday is the worst weekday while Wednesday is the best, closely followed by Thursday.
Jobblue-collar	-0.296	The odds of purchase for blue collar job is 74% that of the baseline category, admin. Blue-collar and housemaid have the lowest expectation of purchases, all else equal, while student and retired “jobs” have the highest.

Task 9 – Investigate a shrinkage method (8 points)

Many candidates struggled to give a clear explanation of how elastic net performs feature selection, only giving some sense of formulas without noting what effect these formulas have.

Candidates sometimes used the wrong set of variables or did not comment on the differences in selected features as asked for.

Elastic net adds to the loglikelihood a penalty based on the magnitude of the estimated coefficients when training the model. The penalty includes both a term based on the sum of the squares of the coefficients, as in ridge regression, and a term based on the sum of the absolute values of the coefficients, as in LASSO regression. An alpha hyperparameter controls how much of each type of term is included, and a lambda hyperparameter controls the size of the overall penalty. The penalty induces shrinkage in the estimated coefficients when they are being optimized, and the inclusion of an absolute value term allows this shrinkage to go all the way to zero, effectively removing the feature.

I use elastic net with alpha equal to 0.5 to create a regularized regression, including the same variables (i.e., age, job, marital, edu_years, housing, loan, phone, month, weekday, PC1, and PC2) used in the full GLM model in Task 6.

Below is the output for the final elastic net regression model using the value of lambda that resulted in the minimum misclassification error:

```
35 x 1 sparse Matrix of class "dgCMatrix"
                           s0
(Intercept)       .          
age               .          
jobblue-collar   -0.164043362
jobentrepreneur   .          
jobhousemaid      .          
jobmanagement     .          
jobretired        0.115439043
jobself-employed  .          
jobservices       .          
jobstudent        0.167404430
jobtechnician     .          
jobunemployed     .          
maritalmarried    .          
maritalsingle     0.003406189
edu_years         .          
housingyes        .          
housingunknown    .          
loanyes           .          
loanunknown       .          
phonelandline    -0.009278094
monthaug         -0.079866517
monthdec          0.060948350
monthjul          0.016881896
monthjun          0.036798320
monthmar          0.372880757
monthmay         -0.749455886
monthnov         -0.214186409
monthoct          0.428990209
monthsep          0.160196267
weekdaymon       -0.025951609
weekdaythu        .          
weekdaytue        .          
weekdaywed        .          
PC1               0.543221403
PC2               .          
```

The AUC were 0.7742 and 0.7659 on the training and test sets respectively.

The elastic net model includes factors from the following features:

- PC1
- Month
- Weekday (only Monday vs. rest of week)
- Job (Only blue-collar, retired, and student vs. all other jobs)

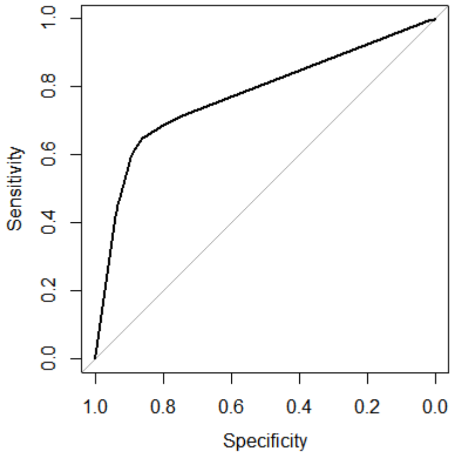
- Marital (Single vs. all other)
- Phone

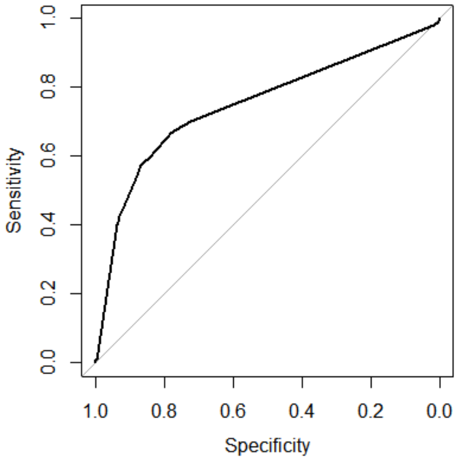
All the selected features using stepwise regression in Task 7 are on the list of variables in the elastic net model. However, due to binarization, not all categories within Weekday and Job are given distinct coefficients as they had been given in the GLM in Task 7. The elastic net model also picked up new variables with Marital and Phone.

Task 10 – Construct a decision tree (6 points)

> *Candidates needed to have considerations related to the use of a decision tree when justifying the variables to be used.*

Using underlying variables instead of the principal components derived from them is helpful here because it is harder to interpret those PCA variables and decision trees are not adversely affected by the presence of highly correlated variables. However, there is no or little information can be gained from splitting on employment after having split on irate, given that "irate" and "employment" are very highly correlated, and the selection of one or the other may be inconsistent depending on the selection of the training data. Therefore, dropping the employment variable is a reasonable choice.





The ROC curves for the train (top) and test (bottom) data are shown above, with fairly similar shapes. The train ROC curve does briefly cross the diagonal line in the upper right, indicating that the model's very lowest probabilities for purchase based on the training data were poor predictions in the test data—these included more purchases than expected, and some overfitting has occurred.

As the AUC measures the area under the ROC curve, a similar story is evident when comparing the train AUC of 0.7744 to the test AUC of 0.7500. The significant reduction in AUC from train to test is a sign that some overfitting has occurred—the ROC curves help to illustrate where that overfitting is happening in this case.

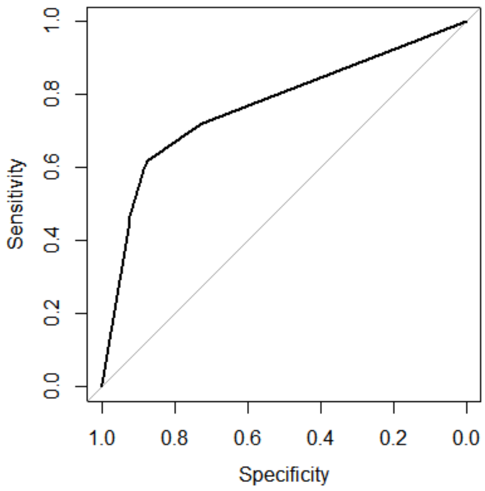
Task 11 – Employ cost-complexity pruning to construct a smaller tree (10 points)

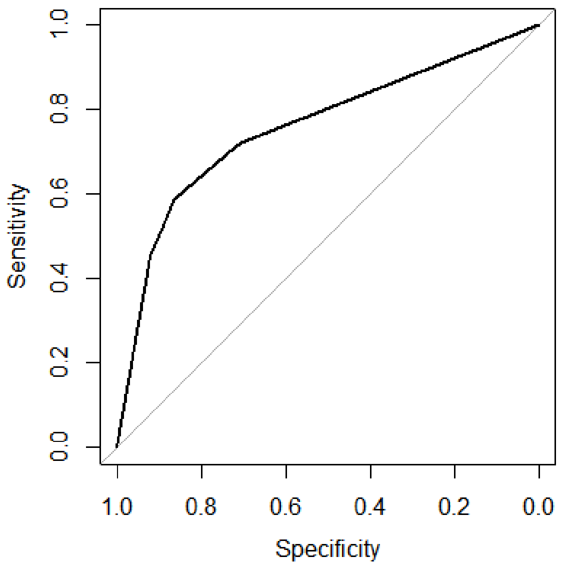
Better candidates explained what the cp table was doing rather than just noting mechanically how to choose the best cp value for pruning. Other valid techniques for pruning the tree than that shown were acceptable.

Better candidates pointed out the reduction in overfitting and interpreted the tree appropriately for marketing rather than just reading out the tree.

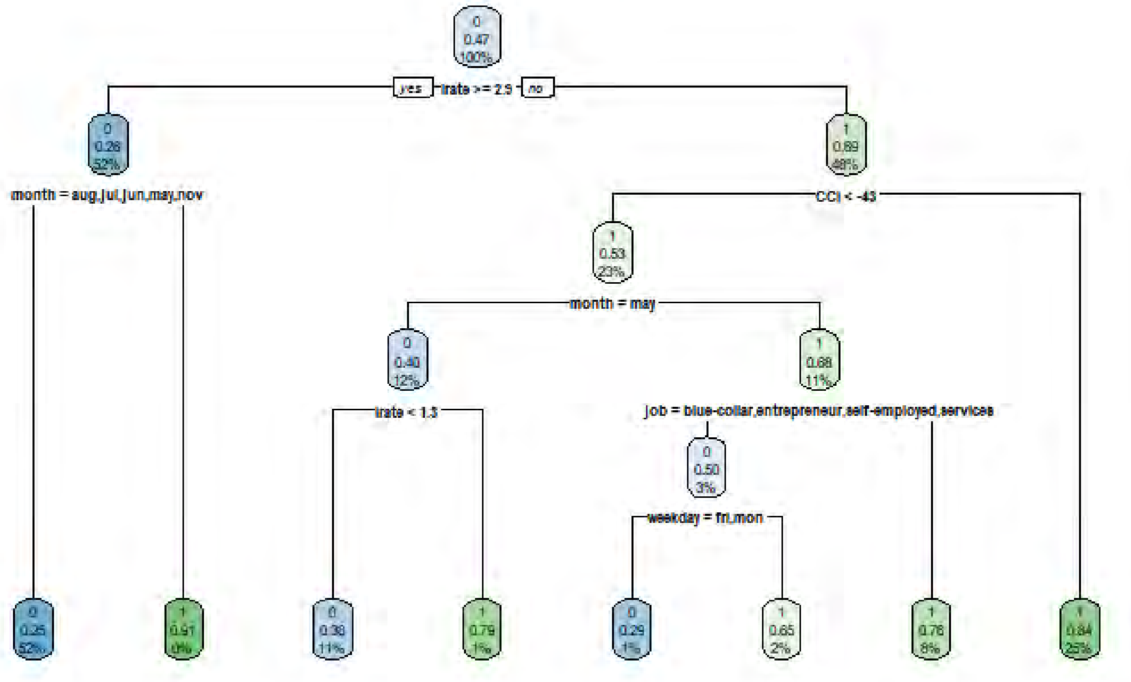
The complexity parameter (CP) is used to find the optimal tree size and reduce the overfitting seen above. The following output is from the initial unpruned tree. The optimal CP is the one that minimizes the cross validation error (in the xerror column). Row 6 accomplishes that, with CP = 0.0021705426. Pruning with this CP value will result in a tree with 7 splits and so 8 leaves.

```
            CP nsplit rel error    xerror       xstd
1 0.3941085271      0 1.0000000 1.0000000 0.01287384
2 0.0263565891      1 0.6058915 0.6058915 0.01161399
3 0.0086821705      3 0.5531783 0.5531783 0.01128533
4 0.0080620155      4 0.5444961 0.5500775 0.01126459
5 0.0058914729      5 0.5364341 0.5364341 0.01117140
6 0.0021705426      7 0.5246512 0.5311628 0.01113454
7 0.0018604651      8 0.5224806 0.5330233 0.01114760
8 0.0008268734     11 0.5168992 0.5320930 0.01114108
9 0.0005000000     16 0.5113178 0.5407752 0.01120139
```





The train (upper) and test (lower) ROC curves are shown above. The test ROC curve no longer hooks upwards at the top right, showing that poor predictions at the lowest probabilities have been addressed. The train and test AUC's are 0.7682 and 0.7593 respectively. While the train AUC had to come down from the 0.7744 of the previous tree due to being a simpler model, the test AUC actually came up from 0.7500 of the previous tree despite the similar model. The overfitting of the prior tree has been reduced with this pruned tree and the simpler model has been shown to be the better predictor on new data.



Of interest are the two leaves that account for the largest proportions of the training data.

- 52% of the past experience used for training falls into the leaf farthest to the left, when interest rates exceed 2.92% in the months from May to August or in November. Only 25% of these made a purchase, and the model predicts no purchase for future prospects in this situation. The combination of high interest rates and summer months appear to be a particularly poor combination for marketing our products, as this group had the lowest historical purchase rate of all eight groups.
- Another 25% of the past experience used for training falls into the leaf farthest to the right, when interest rates are less than 2.92% and the consumer confidence index (CCI) is higher (greater than or equal to -43.5). 84% of these made a purchase, regardless of the month, and the model predicts a purchase for future prospects in this situation. High consumer confidence is a good time to market our products but only when interest rates are low enough.

Task 12 – Choose a model (4 points)

Some candidates did not consider both predictive power and applicability to the business problem, and others gave justifications based on one of these but then chose a model based on the other. This particular business problem did not favor choosing a model solely on AUC given how similar these typically were among models.

Model	Test AUC	Applicability
Full GLM (Task 6)	0.7665	Low: shows trends, but no reduction in variables
GLM StepAIC (Task 8)	0.7669	Medium: Has variable selection, PC1 hard to explain
Elastic Net GLM (Task 9)	0.7500	Medium: Has selection, PC1 hard to explain
Pruned Tree (Task 11)	0.7593	High: straightforward explanations with interactions

To choose a model I will use for advising the marketing department in future campaigns, I consider the balance between predictive power, indicated by a high value of test AUC, and a simpler model for giving more straightforward advice. The table above provides the test AUC's and some considerations for applicability.

I recommend using the pruned decision tree. Among the three GLM models, the one emerging from the stepwise process has both higher predictive power and better applicability as it performs meaningful variable selection while most of the variables are not too hard to explain. However, the principal component variable presents some interpretation difficulties, and this GLM model is hampered by the lack of interaction effects, which may be important. The pruned decision tree is a far simpler model where seemingly important interaction effects are noted. Its predictive power is not far off from that of the more complex GLM, and the ease of explaining this model, without worrying about odds ratios, make this the preferred choice.

Task 13 – Executive summary (20 points)

> *Rather than restating information from prior tasks, candidates were expected to alter their messaging for the intended audience. Often this includes avoiding overly technical language, discussing topics at a different level of detail, and translating performance metrics to be more meaningful to the reader. Brief discussions about approaches attempted are acceptable, but candidates should avoid lengthy discussion about models or techniques that were not ultimately selected. The best candidates were able to incorporate the business context of the problem throughout their summary.*

I have been asked to advise the marketing department at ABC insurance on what efforts will be most productive in terms of purchases for future marketing campaigns for a particular insurance product, based on data collected from a completed marketing campaign. The data has been analyzed using predictive models to bring out what aspects of the marketing campaign have the greatest impacts on whether customers purchase the product. Because the data is specific to just this product, this advice is also specific to just this product. The data did not include any experience in January or February, so no predictions on marketing campaigns for this product in these months can be made.

The data contains 10,000 observations for 14 variables that include personal information about the potential purchaser, the timing of the call, economic indicators at the time of the call, and whether a purchase was made, the variable to be predicted in the future based on the other variables. In this data, 46% of calls resulted in a purchase.

Some records had missing data. Almost 5% of the records did not have the education of the potential purchaser, and because these generally had higher purchase rates, I did not want to remove these records in case other variables helped to explain the higher purchase rates. The average education of other potential purchasers was used as a substitute, but I would like to discuss what leads to missing education data and whether this substitution is appropriate. Where other missing data was encountered, 104 records where dropped where this seemed insignificant to results but in other cases an “unknown” category was created. Modeling proceeded with the remaining 9,896 records.

Some of the economic indicators were highly correlated, so I applied a variety of techniques, from principal components analysis to simply removing one of the variables, employment, to improve the stability of our models.

To test the predictive power of all types of models, the models were trained on 70% of the data and their performance measured only on the remaining 30% of the data not yet seen by the models. The performance metric was “area under the curve”, which describes how accurate the ranking of the probability of purchase is from highest to lowest when compared with purchases made.

During the modeling stage, I examined several options, including a full generalized linear model (GLM) and two reduced versions of this model to isolate the most important factors for distinguishing higher probabilities of purchase. All models in this stage performed reasonably well, but some are simpler than others. I also used a decision tree as it can handle interactions among factors more easily than GLMs, and after it was adjusted to retain only its most significant distinctions among potential purchasers, its results were only slightly less accurate than the GLM models. The advantages of the decision tree outweigh the slight loss of accuracy and is chosen as the model for generating marketing insights.

A summary of the decision tree is provided below in table form:

Proportion of Records	% of Those Purchasing	Interest Rate	Month	Consumer Confidence Index	Job	Weekday
51.7%	25%	Over 2.91%	May, June, July, August, or November	Any	Any	Any
0.5%	91%		October			
11.5%	38%	Under 1.35%	May	Worse than -43.5		
0.7%	79%	Between 1.35% and 2.91%				
1.3%	29%	Under 2.92%	March, April, October, or November		Blue-collar, Entrepreneur, Self-employed, or Services	Monday or Friday
1.8%	65%					Tuesday, Wednesday, or Thursday
7.8%	76%				Admin, Housemaid, Management, Retired, Student, Technician, or Unemployed	Any
24.7%	84%		Any	Better than -43.6	Any	

In this table, the rows on the left pertain to different divisions of the data while the columns show the proportion of records in that division, how many purchased in that division, and what defines the division. The columns are in the order of how often they helped to define a division, and blocks in those columns apply to all rows they cover. Variables not seen were not found to be significant predictors of purchase based on this experience.

The most significant variables are interest rates, month, and consumer confidence index (CCI), and these describe the three quarters of data found in the top and bottom rows of the table. In the top row, with just over half of records, high interest rates during the summer months (or November) led to low proportions of purchases, just 25%, the lowest in the table. While October was a minor exception in this data (with only 0.5% of the records), the general outcome is that it is difficult to sell this product in a high interest rate environment.

In the bottom row, with about a quarter of data, a lower interest rate environment prevailed, and a high 84% purchase rate occurred but only when the CCI was better than -43.6, indicating that both low interest rates and better consumer confidence are needed for a high purchase rate. When consumer confidence was worse in the low interest rate environments, the type of job or even the day of the week could drive significant distinctions in purchase rates. In May, the lowest interest rates drove lower purchases instead of higher, indicating that the direction of interest rates alone is not a reliable predictor.

At a high level, a successful marketing campaign for this product has more to do with market conditions and timing of the calls and less to do with the characteristics of who is called, with only job sometimes generating a distinction. This conclusion is dependent on the data provided and techniques used. I look forward to discussing these results with you in more detail and working together to refine the insights generated thus far.